1. The project that I proposed in Deliverable 1 is to create a classifier that is able to identify the language of the text that is entered into it.

2. I am using a language detection data set from kaggle. There are 17 languages on this data set. https://www.kaggle.com/datasets/basilb2s/language-detection

- To preprocess this data, I made each string of text into lower case and removed any non-letter or non-space characters. I split that data into a training and test set and created a vocabulary of the 10000 most used words from all the text in the training set. To ensure an even distribution of all languages we are testing for in the test set, I let the test set be every 5th data point in the set.
- I then translated all the language identification data into numbers 1-17 and converted the text data into vectors corresponding with the vocabulary.

3. I decided to use the Naive Bayes classifier and at this time it is the model that I am still choosing to use.
    a. I used pandas, re and the Collections library in implementing my model
    b. The data provided was grouped by language (all english points first, french second etc.) so to split my data set I removed every 5th data point from the set and added that to my test set. The main hyperparameter decision is in my model was the size of the vocabulary. I let that be 10000 words as this seemed like an adequately large number of words to pick out the most frequently used word in each desired language without overfitting. I may need to adjust this to pull the 300-400 most used words in each language, as the data set does contain more english text data so my current vocabulary may be overwhelmed with english words.
    c. To test my model I input test vectors into a naive bayes function that I wrote and verify the output of this function against the expected result. At this time my model is not fully working and I was not able to debug it before reading week. I will need to figure out this issue with the naive bayes function as it classifying all inputs as english.
    d. As stated above I am working now on solving the issues with my model.
4. As stated in question 3 my model is not yet functional. I will soon be testing it against the expected outcomes of the test set.
5. My next steps are to debug the model, aiming for at least 70% accuracy in language detection. Afterthat I will look at adjusting the vocabulary I use in this model to maybe adjust the size or implement other measures to ensure there is an even distribution of words from each language I am looking to identify represented.